Conformal clustering for functional variables,
With application to electricity consumption curves

Ilia Nouretdinov, Matteo Fontana, James Gammerman, Laura Shemilt, Daljit Rehal
CLRC, Royal Holloway University of London; Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Italy; Centrica, UK

Abstract

Conformal Clustering (CC) is a clustering technique which also allows for anomaly detection. It involves finding a so-called 'region of conformity' in the feature space: at a pre-selected significance level ε, the data points within this region are grouped into clusters, while all data points outside it are considered anomalies.

In the existing literature on CC, data has typically been considered as finite-dimensional feature vectors. However, much existing data is found in functional form - for example time series data. In this work we generalise the CC technique to the domain of functional data. More specifically, we use CC to clean data from energy consumption curves, with the aim of subsequently disaggregating these energy curves into their components.

Our experiments play two roles. Firstly, to validate our expectation that conformal clustering is an effective technique for data cleaning. And secondly, to confirm that a functional approach to data analysis can provide new insights that are lacking in a vector approach.

Functional Conformal Clustering Algorithm

INPUT: significance level ε
INPUT: dataset in functional form

- Set three auxiliary functions:

1) A projection function P for dimensionality reduction + a grid G in the projection space
2) A NCM function A for measuring strangeness of the functions
3) An enveloping (smoothing) function E for de-projection

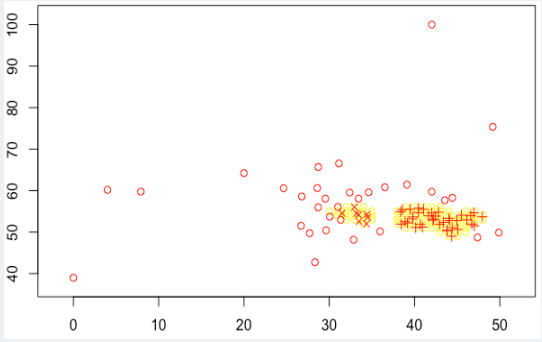
- Apply projection function P to the data to get the projection.
- Test each point of the grid G for conformity with the input data set, using the NCM function A, and output p-value.
- Find the prediction set R of grid points with p-value over ε
- Represent R as a union of connected components.
- Apply E to the components to find their envelopes in the projection space
- OUTPUT their pre-images in the original data space.

Functional Conformal Clustering Example

- We used a sample of the publicly available Pecan Street dataset, giving energy consumption of many households in Texas, USA.
- Data includes hourly energy consumption both in total and broken down into 66 measures of consumption (either rooms or appliances)
- For each household, we create an averaged daily consumption profile. At any time point h= 0,1,2,..,H–1 hours with H= 24, we calculate average consumption $f_0(h)$ at this time point over all days
- Before applying Conformal Clustering, the data is pre-processed:

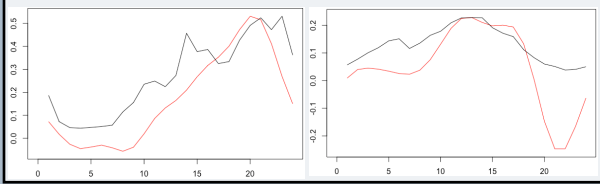
1. Fourier smoothing converts each of the profiles to functional form
2. An optional functional transformation: including the derivative.

When the algorithm is applied, two-dimensional PCA is used as the projection function, and NCM is based on 5-nearest neighbours.
The plot axes are two PCA components:



- Aim of the clustering: disaggregation of the profile into its components (such as room or device).
- Disaggregation is done by PCA decomposition, which can be improved by using the examples from the largest cluster only.
- To evaluate the results of disaggregation, we compare the obtained principal components with true individual average profiles for appliances/rooms, by cosine similarity.

Examples below: component shown in black and its most similar appliance in red.
PCA comp.1 - light-plugs-3 (similarity 0.912);
PCA comp.2 - sprinkler (similarity 0.658)
X-axis = time (hours), y-axis = energy consumption (up to proportionality)



Results & Conclusions

Smoothing	-	-	$K=5$	$K=10$	$K=5$	$K=10$
Derivative	-	-	-	-	+	+
CC	-	Cleaned at $\varepsilon = 0.5$				
No. of exa.:	77	40	39	35	37	45
PCA1	0.992	0.992	0.898	0.892	0.912	0.959
PCA2	0.559	0.495	0.800	0.745	0.658	0.534
PCA3	0.395	0.467	0.552	0.621	0.640	0.639
PCA4	0.479	0.529	0.623	0.636	0.590	0.446
PCA5	0.231	0.479	0.477	0.550	0.505	0.357
PCA6	0.340	0.256	0.198	0.186	0.212	0.431
PCA7	0.254	0.208	0.236	0.262	0.199	0.233
PCA8	0.296	0.094	0.100	0.101	0.110	0.186
PCA9	0.215	0.122	0.107	0.113	0.144	0.138
PCA10	0.332	0.275	0.189	0.312	0.149	0.174
med.(1–5)	0.479	0.495	0.623	0.636	**0.640**	0.534
med.(1–9)	0.340	0.467	0.477	**0.550**	0.505	0.431

1. Conformal Clustering improves the quality of disaggregation.

2. The advantage of functional data analysis as opposed to vector data analysis: Fourier smoothing is an important means of increasing the quality: from 0.495 to 0.534–0.640, or from 0.467 to 0.431–0.550.

3. With regards to the level of Fourier smoothing, K = 10 is preferable to K = 5, but there is little difference

4. For the top five principal components, the best result is achieved when Fourier smoothing with K= 5 is combined with using the derivative; this shows that taking the derivative is a useful option within the functional data analysis approach.

5. In almost all the cases, deleting the small clusters and outliers show better results than deleting only the outliers (omitted from table).

We can conclude that a functional interpretation of data gives new possibilities in data analysis that are not achievable in vector approach, and that conformal clustering is a useful pre-processing technique.

Contact information and acknowledgements

Corresponding author:
Ilia Nouretdinov
i.r.nouretdinov@rhul.ac.uk

This work was funded by Centrica plc, AstraZeneca, and Mittie.